

Deciphering Developmental Disorders Study, (2017) Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, 542, pp. 433-438. (doi:[10.1038/nature21062](https://doi.org/10.1038/nature21062))

This is the author's final accepted version.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/169399/>

Deposited on: 21 September 2018

# The prevalence and architecture of de novo mutations in dominant developmental disorders

The Deciphering Developmental Disorders Study

## Acknowledgments

We thank the families for their participation and patience. We are grateful to the Exome Aggregation Consortium for making their data available. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the UK Department of Health, and the Wellcome Trust Sanger Institute (grant WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the UK Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). The research team acknowledges the support of the National Institutes for Health Research, through the Comprehensive Clinical Research Network. The authors wish to thank the Sanger Human Genome Informatics team, the DNA pipelines team and the Core Sequencing team for their support in generating and processing the data. D.R.F. is funded through an MRC Human Genetics Unit program grant to the University of Edinburgh. Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England ([www.nhsbt.nhs.uk](http://www.nhsbt.nhs.uk)), which has supported field work and other elements of the trial. DNA extraction and genotyping was funded by the National Institute of Health Research (NIHR), the NIHR BioResource (<http://bioresource.nihr.ac.uk/>) and the NIHR Cambridge Biomedical Research Centre ([www.cambridge-brc.org.uk](http://www.cambridge-brc.org.uk)). The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), and NIHR Research Cambridge Biomedical Research Centre. A complete list of the investigators and contributors to the INTERVAL trial is provided in Moore et al. (2014).

## Abbreviations

PTV: Protein-Truncating Variant

SNV: Single Nucleotide Variant

DNM: De Novo Mutation

## Author Contributions

Baralle, D., Temple, I.K., et al. The Deciphering Developmental Disorders Study

## Introductory text

Children with severe, undiagnosed developmental disorders (DDs) are enriched for damaging de novo mutations (DNMs) in developmentally important genes. We exome sequenced 4,294 families with children with DDs, and meta-analysed these data with published data on 3,287 children with similar disorders. We show that the most significant factors influencing the diagnostic yield of de novo mutations are the sex of the child, the relatedness of their parents and the age of both father and mother. We identified 95 genes enriched for damaging de novo mutation at genome-wide significance ( $P < 5 \times 10^{-7}$ ), including fourteen genes for which compelling data for causation was previously lacking. The large number of genome-wide significant findings allow us to demonstrate that, at current cost differentials, exome sequencing has much greater power than genome sequencing for novel gene discovery in genetically heterogeneous disorders. We estimate that 42.5% of our cohort likely carry pathogenic de novo single nucleotide variants (SNVs) and indels in coding sequences, with approximately half operating by a loss-of-function mechanism, and the remainder being gain-of-function. We established that most haploinsufficient developmental disorders have already been identified, but that many gain-of-function disorders remain to be discovered. Extrapolating from the DDD cohort to the general population, we estimate that de novo dominant developmental disorders have an average birth prevalence of 1 in 168 to 1 in 377, depending on parental age.

## Main text

Within the Deciphering Developmental Disorders study<sup>1</sup> we recruited 4,294 individuals with severe undiagnosed developmental disorders, most of which were the only affected family member. We sequenced the exomes of these individuals and their parents. Analyses of 1,133 of these trios were described previously<sup>1</sup>. We generated a high sensitivity set of 8,361 candidate DNMs in coding or splicing sequence (mean of 1.95 DNMs per proband), while removing systematic erroneous calls (Supplementary Table 1). 1,624 genes contained two or more DNMs in unrelated individuals.

### **Figure 1 - factors influencing presence of pathogenic DNM**

Twenty-three percent of individuals had likely pathogenic protein-truncating or missense DNMs within a clinically curated set of genes robustly associated with developmental disorders (<http://www.ebi.ac.uk/gene2phenotype>). We investigated factors associated with whether an individual had a likely pathogenic DNM in these curated genes (Figure 1 A, B). We observed that males were less likely to carry a likely pathogenic DNM ( $P = 1.8 \times 10^{-4}$ ; OR 0.75, 0.65 - 0.87 95% CI), as has also been observed in autism<sup>2</sup>. We also observed increased likelihood of having a pathogenic DNM with the extent of speech delay ( $P = 0.00123$ ), but not other indicators of severity. Furthermore, the total genomic extent of autozygosity (due to parental relatedness) was negatively correlated with the likelihood of having a pathogenic DNM ( $P = 1.4 \times 10^{-5}$ ), for every  $\log_{10}$  increase in autozygous length, the probability of having a pathogenic DNM dropped by 7.5%, likely due to increasing burden of recessive causation (Figure 1 C). Nonetheless, 6% of individuals with autozygosity equivalent to a first cousin union or greater had a likely pathogenic DNM, underscoring the importance of considering de novo causation in all families.

Paternal age has been shown to be the primary factor influencing the number of DNMs in a child<sup>3,4</sup>, and thus is expected to be a risk factor for pathogenic DNMs. While paternal age was only weakly associated with likelihood of having a pathogenic DNM ( $P = 0.016$ ), focusing on the minority of DNMs that were truncating and missense variants in known DD-associated genes limits our power to detect such an effect. Analysing all 8,409 high confidence exonic and intronic autosomal DNMs confirmed a strong paternal age effect ( $P = 1.4 \times 10^{-10}$ , 1.53 DNMs/year, 1.07-2.01 95% CI), as well as highlighting a weaker, independent, maternal age effect ( $P = 0.0019$ , 0.86 DNMs/year, 0.32-1.40 95% CI, Figure 1 D, E), as has recently been described in whole genome analyses<sup>5</sup>.

**Figure 2 – Manhattan plot**

**Figure 3 – phenotypic summary of ‘previously not compelling’ genes**

**Table 1 – summary of DNMs and  $p$  values in ‘previously not compelling’ genes**

We identified genes significantly enriched for damaging DNMs by comparing the observed gene-wise DNM count to that expected under a null mutation model<sup>6</sup>, as described previously<sup>1</sup>. We meta-analysed with 4,224 published DNMs in 3,287 affected individuals from thirteen exome or genome sequencing studies (Supplementary Table 3)<sup>7-18</sup> that exhibited a similar excess of DNMs in our curated set of DD-associated genes (Supplementary Figure 1). We found 93 genes with genome-wide significance ( $P < 5 \times 10^{-7}$ ), 76 of which were in our curated gene set (Supplementary Table 4). Some disorders are considerably more clinically distinctive than others (Supplementary Figure 2). To increase power to detect novel DD-associated genes, we then excluded individuals with likely pathogenic variants in known DD-associated genes<sup>1</sup>, leaving 3,158 probands from our cohort, along with 2,955 probands from the meta-analysis studies (Supplementary Table 5). In this subset, we identified fourteen genome-wide significant genes for which compelling evidence for causation has not previously been presented: *CDK13*, *CHD4*, *CNOT3*, *CSNK2A1*, *GNAI1*, *KCNQ3*, *MSL3*, *PPM1D*, *PUF60*, *QRICH1*, *SET*, *SUV420H1*, *TCF20*, and *ZBTB18* ( $P < 5 \times 10^{-7}$ , Table 1, Supplementary Figure 3). The clinical features associated with these novel disorders are summarised in Figure 3. *QRICH1* and *SET* would not achieve genome-wide significance without excluding individuals with likely pathogenic variants. We found *USP9X* and *ZC4H4* had a genome-wide significant excess in female probands, indicating these genes have X-linked dominant modes of inheritance in addition to previously reported X-linked recessive mode of inheritance in males. In addition, we identified a novel seizure disorder caused by truncating mutations in *SMC1A* ( $P = 6.5 \times 10^{-19}$ ), a DD-associated gene in which missense mutations cause a distinct disorder, Cornelia de Lange Syndrome. Only one PTV mutation in *SMC1A* had previously been reported<sup>19</sup>.

We additionally exome sequenced 566 ‘case’ individuals with DDs for which parental DNA was not available, and 4,100 ‘control’ individuals without known neurodevelopmental phenotypes<sup>20</sup>. Cases exhibited an excess of rare PTVs in DD-associated dominant genes ( $P = 2.7 \times 10^{-22}$ ; OR 4.78, 4.07 - 5.63 95% CI). After excluding 90 cases with rare PTVs in dominant DD-associated genes, cases still had a modest exome-wide excess of rare PTVs ( $P = 4.3 \times 10^{-3}$ ; OR 1.07, 1.04 - 1.09 95% CI). Furthermore, we found a significant enrichment of rare PTVs in the fifteen novel developmental disorder genes ( $P = 1.5 \times 10^{-5}$ ; OR = 11.7, 6.5 - 21.2 95% CI), providing additional evidence that disruptive variants in these newly associated genes confer risk for developmental disorders.

The above analyses focus exclusively on the genetic evidence for association. We explored two alternative strategies for integrating phenotypic data: statistical assessment of phenotypic similarity between individuals sharing candidate DNMs in the same gene (as we described previously<sup>21</sup>) and phenotypic stratification. We found that while combining genetic evidence and phenotypic similarity increased the significance of some known DD-associated genes considerably, significance decreased for a larger number of DD-associated genes that cause relatively indistinct disorders (Supplementary Figure 4 A). Therefore, we did not incorporate phenotypic similarity in the gene discovery analyses described above.

We also investigated phenotypic stratification by comparing gene-wise analyses of the 20% of individuals who had experienced seizures, with gene-wise analyses of the entire cohort, to see if it increased power to detect known seizure-associated genes (Supplementary Figure 4 B). Fifteen seizure-associated genes were genome-wide significant within both the seizure-only and the entire-cohort analyses. Furthermore, nine seizure-associated genes were genome-wide significant in the entire cohort but not in the seizure subset. Of the 285 individuals with truncating or missense DNMs in known seizure-associated genes, 56% of individuals had not experienced seizures. This observation is not likely to be due to the individuals manifesting seizures having more damaging DNMs as the proportions of truncating mutations were not significantly different between individuals with and without seizures ( $P = 0.05$ ). These findings suggest that there is sufficient shared genetic etiology between individuals with seizures and individuals with other neurodevelopmental disorders in our cohort that increased sample size far outweighs greater phenotypic specificity.

#### ***Figure 4 - power of exome and genome sequencing under different assumptions***

The large number of genome-wide significant genes identified in the analyses above allows us to compare empirically different experimental strategies for novel gene discovery in a genetically heterogeneous cohort such as ours. We compared the power of exome and genome sequencing to detect genome-wide significant genes, assuming that budget and not samples are limiting, under different scenarios of cost ratios and sensitivity ratios (Figure 4). We found that at current cost ratios (exome costs 30-40% of a genome) and with a plausible sensitivity differential (genome detects 5% more exonic variants than exome<sup>22</sup>) exome sequencing detects more than twice as many genome-wide significant genes. These empirical estimates were consistent with power simulations for identifying dominant loss-of-function genes (Supplementary Figure 5). The close agreement of empirical estimates and power simulations based on germline mutation rates are consistent with few *de novo* mutations being lost due to prenatal lethality. In summary, while genome sequencing gives greatest sensitivity to detect pathogenic variation in a single individual, exome sequencing is more powerful for novel gene discovery (and, analogously, likely delivers lower cost per diagnosis).

#### ***Figure 5 – DNM excess for recognisability and consequence, also by constraint quantile***

Our previous simulations suggested that analysis of a cohort of 4,294 DDD families ought to be able to detect approximately half of all haploinsufficient DD-associated genes at genome-

wide significance<sup>1</sup>. Empirically, we identified 47% (50/107) of haploinsufficient genes previously robustly associated with neurodevelopmental disorders. We hypothesised that genetic testing prior to recruitment into our study may have depleted the cohort of the most clinically recognisable disorders. Indeed, we observed that the genes associated with the most clinically recognisable disorders were associated with a significant, three-fold lower enrichment of truncating DNMs than other DD-associated genes (~40X enrichment vs ~120X enrichment, Figure 5 A). Removing these most recognisable disorders from the analysis, we identified 55% (42/76) of the remaining haploinsufficient DD-associated genes. The known DD-associated haploinsufficient genes that did not reach genome-wide significance were clearly enriched for those with lower mutability, which we would expect to lower power to detect in our analyses. We identified DD-associated genes (e.g. NRXN2) with high mutability, low clinical recognisability and yet no signal of enrichment for DNMs in our cohort (Supplementary Figure 6). Our analyses call into question whether these genes really are associated with haploinsufficient neurodevelopmental disorders and highlights the potential for well-powered gene discovery analyses to refute prior credence regarding gene causation.

We estimated the likely prevalence of pathogenic missense and truncating DNMs within our cohort by increasing the stringency of called DNMs until the observed synonymous DNMs equated to that expected under the null mutation model, and then quantifying the excess of observed missense and truncating DNMs (Figure 5 B). We observed an excess of 591 truncating and 1,236 missense mutations, suggesting 42.5% of the cohort has a pathogenic DNM. The vast majority of synonymous DNMs are likely to be benign, as evidenced by them being distributed uniformly (Figure 5 C) among genes irrespective of their tolerance of truncating variation in the general population (as quantified by the probability of being LoF-intolerant (pLI) metric<sup>23</sup>). By contrast, missense and truncating DNMs are significantly enriched in genes with the highest probabilities of being intolerant of truncating variation (Figure 5 C). Only 51% (923/1816) of these excess missense and truncating DNMs mutated DD-associated dominant genes, with the remainder likely to affect genes not yet associated with DDs. A much high proportion of the excess truncating DNMs (70%) than missense DNMs (42%) mutated known DD-associated genes, suggesting that whereas most haploinsufficient DD-associated genes have already been identified, there remain to be discovered many DD-associated genes characterised by pathogenic missense DNMs.

We sought to estimate the relative proportion of gain-of-function and loss-of-function missense DNMs in our cohort, taking advantage of the different population genetic characteristics of known gain-of-function and loss-of-function DD-associated genes. Specifically, we observed that, as might be expected, these two classes of DD-associated genes are differentially depleted of truncating variation in the general population (pLI metric<sup>23</sup>). We modelled the observed pLI distribution of excess missense DNMs as a mixture of the pLI distributions of known gain-of-function and loss-of-function DD-associated genes (Figure 5 D, E), and estimated that 63% of excess missense DNMs are likely gain-of-function. If we assume that all truncating mutations are operating by a loss-of-function mechanism, then 57% of excess missense and truncating DNMs are loss-of-function and 43% are gain-of-function.

**Figure 6- Parental age vs birth prevalence, underplotted with parental age distributions**

We estimated the birth prevalence of dominant developmental disorders by using the germline mutation model to calculate the expected cumulative germline mutation rate of truncating DNMs in haploinsufficient DD-associated genes and scaling this upwards based on the composition of excess DNMs in the DDD cohort described above (see Methods), correcting for disorders that are under-represented in our cohort as a result of prior genetic testing (e.g. clinically-recognisable disorders and large pathogenic CNVs identified by chromosomal microarray analysis). This gives a mean prevalence estimate of 0.42%, or 1 in 235 births. By factoring in the paternal and maternal age effects on the mutation rate (Figure 1) we modelled age-specific estimates of birth prevalence (Figure 6) that range from 1 in 377 (both mother and father aged 20) to 1 in 168 (mother and father aged 45).

In summary, we have shown that dominant mutations account for approximate half of the genetic architecture of severe developmental disorders, and are split roughly equally between loss-of-function and gain-of-function. Whereas most haploinsufficient DD-associated genes have already been identified, currently many activating and dominant negative DD-associated genes have eluded discovery. This elusiveness likely results from these disorders being individually rarer, as a result of being caused by a relatively small number of missense mutations within each gene. We have assessed empirically different experimental and analytical strategies for identifying DD-associated genes. Discovery of the remaining dominant developmental disorders will be driven by larger studies and novel, more powerful, analytical strategies for disease-gene association that leverage gene-specific patterns of population variation, specifically the observed depletion of damaging variation. We have estimated the mean birth prevalence of dominant developmental disorders to be 1 in 236, which is greater than the combined impact of the three aneuploidies (Down, Edwards, Patau)<sup>24</sup> and highlights the cumulative population morbidity and mortality imposed by these disorders.



## Methods

Being drafted.

## Figures

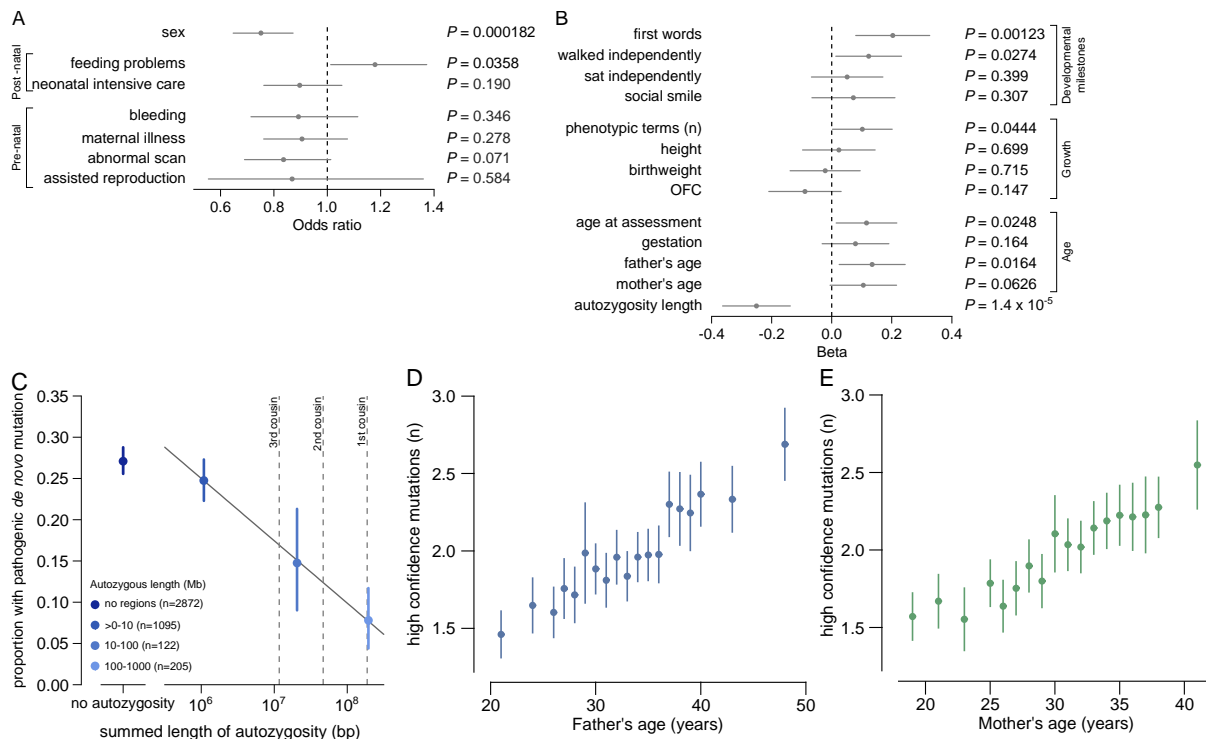


Figure 1: Association of phenotypes with presence of likely pathogenic *de novo* mutations. A) Odds ratios and 95% confidence intervals (CI) for binary phenotypes. Positive odds ratios are associated with increased risk of pathogenic *de novo* mutations when the phenotype is present. P-values are given for a Fisher's Exact test. B) Beta coefficients and 95% CI from logistic regression of quantitative phenotypes versus presence of a pathogenic *de novo* mutation. All phenotypes aside from length of autozygous regions were corrected for gender as a covariate. The developmental milestones (age to achieve first words, walk independently, sit independently and social smile) were log-scaled before regression. The growth parameters (height, birthweight and OFC) were evaluated as absolute distance from the median. C) Relationship between length of autozygous regions chance of having a pathogenic *de novo* mutation. The regression line is plotted as the dark gray line. The 95% confidence interval for the regression is shaded gray. The autozygosity lengths expected under different degrees of consanguineous unions are shown as vertical dashed lines. n, number of probands in each autozygosity group. D) Relationship between age of fathers at probands birth and number of high confidence *de novo* mutations. n, number of high confidence *de novo* mutations. E) Relationship between age of mothers at probands birth and number of high confidence *de novo* mutations. n, number of high confidence *de novo* mutations.

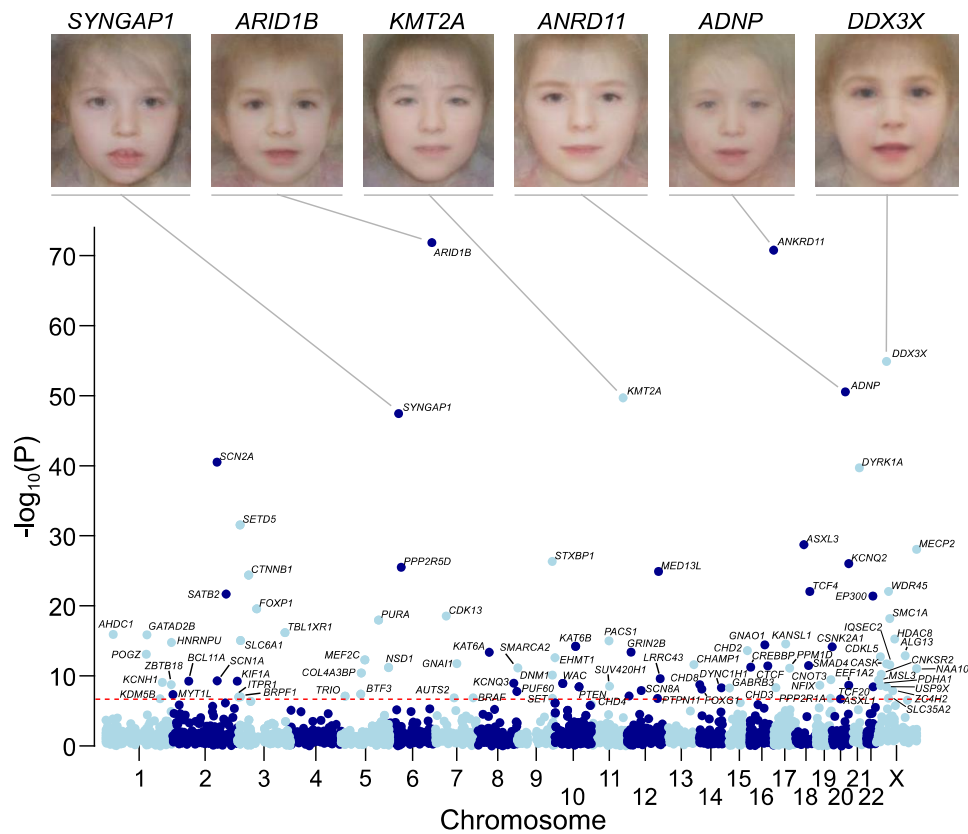


Figure 2: Initial genes exceeding genome-wide significance. Manhattan plot of combined  $P$ -values across all tested genes. The red dashed line indicates the threshold for genome-wide significance ( $P < 5 \times 10^{-7}$ ). Genes exceeding this threshold have HGNC symbols labelled. Composite facial images from individuals with DNMs in selected genes are included for the six most-significantly associated genes.

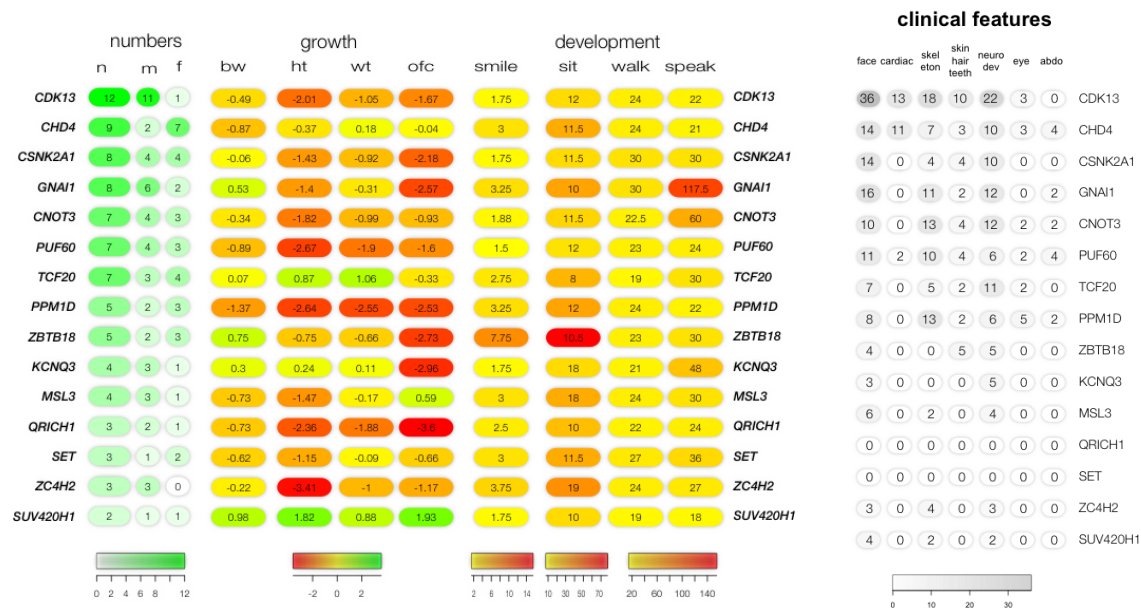


Figure 3: Phenotypic summary of genes without previous compelling evidence. Phenotypes are grouped by type. The first group indicates counts of individuals with DNMs per gene by sex. The second group indicates mean values for growth parameters: birthweight (bw), height (ht), weight (wt), occipitofrontal circumference (ofc). Values are given as standard deviations from the healthy population mean. The third group indicates the mean age for achieving developmental milestones: age of first social smile, age of first sitting unassisted, age of first walking unassisted and age of first speaking. Values are given in months. The final group summarises Human Phenotype Ontology (HPO)-coded phenotypes per gene, as counts of HPO-terms within different clinical categories.

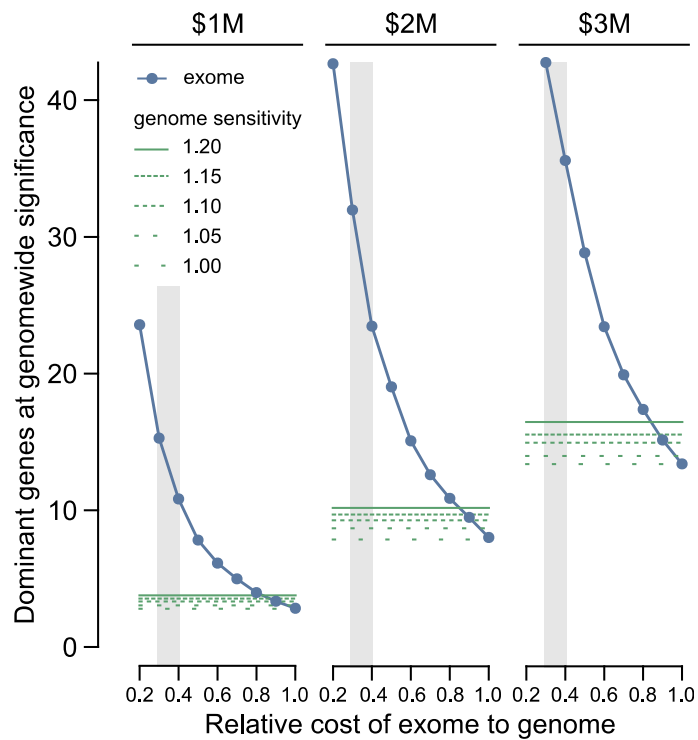
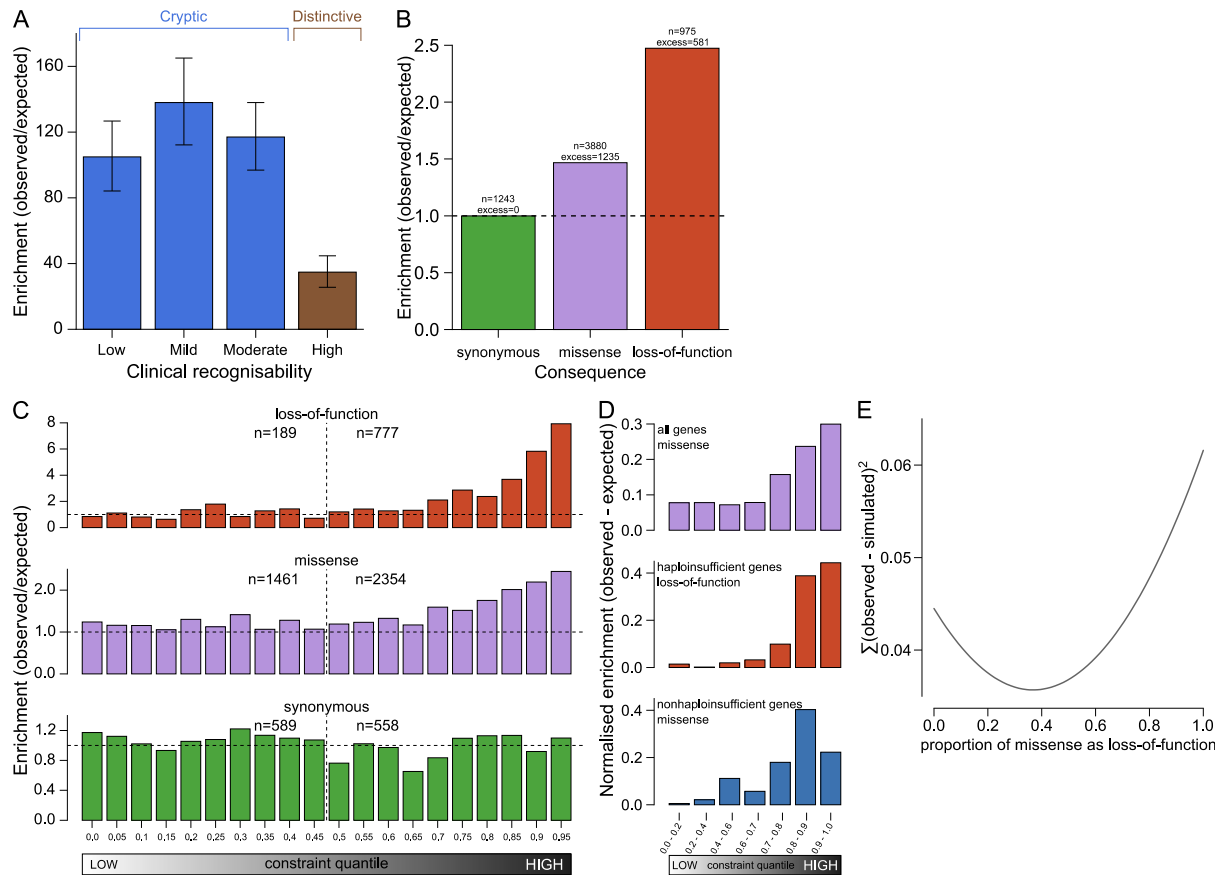


Figure 4: Power of genome versus exome sequencing to discover novel genes. The regions where exome sequencing costs 30-40% of genome sequencing are shaded with a gray background, which corresponds to the price differential in 2015.



**Figure 5: Excess of *de novo* mutations (DNMs).** A) Enrichment ratios of observed to expected loss-of-function DNMs by clinical recognisability for dominant haploinsufficient neurodevelopmental genes. B) Enrichment of DNMs by consequence. C) Enrichment ratios of observed to expected DNMs by constraint quantile for loss-of-function, missense and synonymous DNMs. Counts of DNMs in each lower and upper half of the quantiles are provided. D) Normalised excess of observed to expected DNMs by constraint quantile. This includes missense DNMs within all genes, loss-of-function and missense DNMs in dominant haploinsufficient genes and missense DNMs in dominant nonhaploinsufficient genes (genes with dominant negative or activating mechanisms). E) Goodness-of-fit for mixture models across the range of loss-of-function proportions.

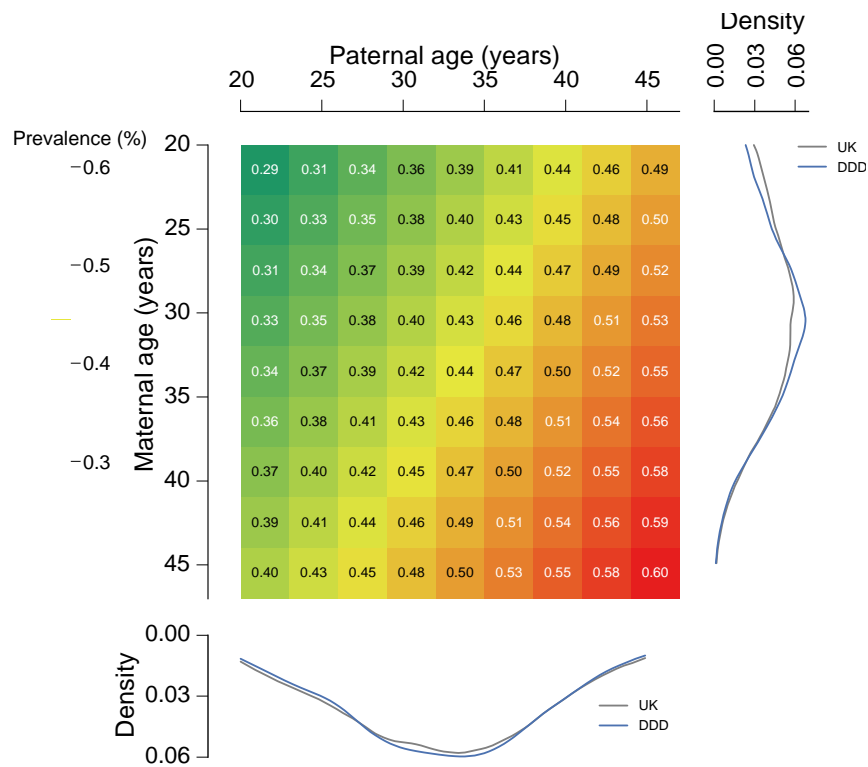


Figure 6: Prevalence of live births with developmental disorders caused by dominant *de novo* mutations (DNMs). The prevalence within the general population is provided as percentage for combinations of parental ages, extrapolated from the maternal and paternal rates of DNMs. Distributions of parental ages within the DDD cohort and the UK population are shown at the matching parental axis.

## Tables

Table 1: Genes achieving genome-wide significant statistical evidence without previous compelling evidence for being developmental disorder genes. The numbers of unrelated individuals with independent *de novo* mutations are given for protein truncating variants (PTV) and missense variants. If any additional individuals were in other cohorts, that number is given in brackets. The *P*-value reported is the minimum *P*-value from the testing of the DDD dataset or the meta-analysis dataset. The subset providing the *P*-value is also listed. Mutations are considered clustered if the *P*-value proximity clustering of *de novo* mutations is less than 0.01.

Gene	Missense	PTV	P-value	Test	Clustering
<i>CDK13</i>	10	1	$3.2 \times 10^{-19}$	DDD	Yes
<i>GNAI1</i>	7 (1)	1	$2.1 \times 10^{-13}$	DDD	No
<i>CSNK2A1</i>	7	0	$1.4 \times 10^{-12}$	DDD	Yes
<i>PPM1D</i>	0	5 (1)	$6.3 \times 10^{-12}$	Meta	No
<i>CNOT3</i>	5	2 (1)	$5.2 \times 10^{-11}$	DDD	Yes
<i>MSL3</i>	0	4	$2.2 \times 10^{-10}$	DDD	No
<i>KCNQ3</i>	4 (3)	0	$3.4 \times 10^{-10}$	Meta	Yes
<i>ZBTB18</i>	1 (1)	4	$1.4 \times 10^{-9}$	DDD	No
<i>PUF60</i>	4 (1)	3	$2.6 \times 10^{-9}$	DDD	No
<i>TCF20</i>	1	5	$2.7 \times 10^{-9}$	DDD	No
<i>SUV420H1</i>	0 (2)	2 (3)	$2.9 \times 10^{-9}$	Meta	No
<i>CHD4</i>	8 (1)	1	$7.6 \times 10^{-9}$	DDD	No
<i>SET</i>	0	3	$1.2 \times 10^{-7}$	DDD	No
<i>QRICH1</i>	0	3 (1)	$3.6 \times 10^{-7}$	Meta	No



## References

1. The Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-228 (2015).
2. Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am J Hum Genet* **94**, 415-25 (2014).
3. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).
4. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat Genet* **48**, 126-33 (2016).
5. Wong, W.S. *et al.* New observations on maternal age effect on germline de novo mutations. *Nat Commun* **7**, 10486 (2016).
6. Samocha, K.E. *et al.* A framework for the interpretation of de novo variation in human disease. *Nature Genetics* **46**, 944-950 (2014).
7. De Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *The New England Journal of Medicine* **367**, 1921-9 (2012).
8. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285-299 (2012).
9. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 1-7 (2012).
10. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674-82 (2012).
11. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
12. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-3 (2013).
13. Epi4K Consortium & Epilepsy Phenome/Genome Project. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-21 (2013).
14. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).
15. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project & Epi4K Consortium. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am J Hum Genet* **95**, 360-70 (2014).
16. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).
17. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).
18. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344-7 (2014).
19. Lebrun, N. *et al.* Early-onset encephalopathy with epilepsy associated with a novel splice site mutation in SMC1A. *American journal of medical genetics. Part A* (2015).
20. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).

21. Akawi, N. *et al.* Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature Genetics* **47**, 1363-1369 (2015).
22. Meynert, A.M., Ansari, M., FitzPatrick, D.R. & Taylor, M.S. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15**, 247 (2014).
23. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* **X**, XX-XX (2015).
24. Springett, A. *et al.* Congenital Anomaly Statistics 2011: England and Wales. (2013).

## Supplementary tables

Table provided in external spreadsheet.

Supplementary Table 1: Table of *de novo* mutations in the 4,294 DDD individuals. The table includes sex, chromosome, position, reference and alternate alleles, HGNC symbols, VEP consequences, and validation status where available. Individual IDs are available on request. This list excludes the sites that failed validations, but includes sites that passed validation (confirmed), sites that were uncertain (uncertain), and sites that were not tested by secondary validation (NA). Genome positions are given as GRCh37 coordinates.

Supplementary Table 2: Proportion of DDD cohort with phenotypic terms that relate to the disorders included in the external cohorts in the meta-analyses.

Disorder	Root terms	Proportion
Autism spectrum disorder	HP:0000729	0.114
Congenital heart disorder	HP:0002564	0.106
Intellectual disability	HP:0001249,HP:0012443,HP:0100543	0.817
Schizophrenia	HP:0100753	0.000
Seizures	HP:0001250	0.199

Supplementary Table 3: Details of cohorts used in meta-analyses. This includes numbers of individuals by sex and publication details.

Phenotype	Year	Male	Female	Note	Citation
Intellectual disability	2012	47	53		De Ligt, et al. <sup>7</sup>
Autism spectrum disorder	2012	314	29	subset of lossifov, et al. <sup>14</sup>	lossifov, et al. <sup>8</sup>
Autism spectrum disorder	2012	151	58	subset of lossifov, et al. <sup>8</sup>	O’Roak, et al. <sup>9</sup>
Intellectual disability	2012	19	32		Rauch, et al. <sup>10</sup>
Autism spectrum disorder	2012	157	68	subset of lossifov, et al. <sup>14</sup>	Sanders, et al. <sup>11</sup>
Seizures	2013	156	108	subset of EuroEPINOMICS-RES Consortium, et al. <sup>15</sup>	Epi4K Consortium and Epilepsy Phenome/Genome Project <sup>13</sup>
Congenital heart disease	2013	220	142		Zaidi, et al. <sup>12</sup>
Seizures	2014	54	38		EuroEPINOMICS-RES Consortium, et al. <sup>15</sup>
Schizophrenia	2014	308	317		Fromer, et al. <sup>17</sup>
Intellectual disability	2014	0	0	subset of De Ligt, et al. <sup>7</sup>	Gilissen, et al. <sup>18</sup>
Autism spectrum disorder (normal IQ)	2014	1099	74	Counts are for individuals with IQ >= 70.	lossifov, et al. <sup>14</sup>
Autism spectrum disorder	2014	446	112	Probands with IQ < 70.	lossifov, et al. <sup>14</sup>
Autism spectrum disorder	2014	1192	253	Counts are extrapolated from the sex ratio of individuals with <i>de novos</i> .	De Rubeis, et al. <sup>16</sup>

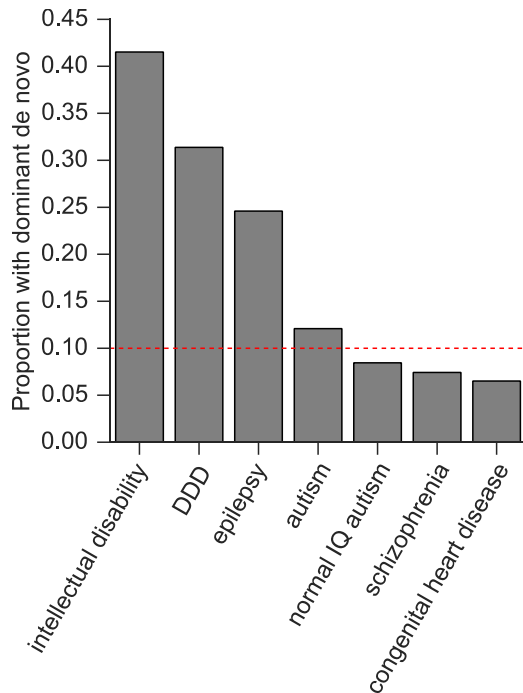
Table provided in external spreadsheet.

Supplementary Table 4: Genes with genome-wide significant statistical evidence to be developmental disorder genes. The numbers of unrelated individuals with independent *de novo* mutations are given for protein truncating variants (PTV) and missense variants. If any additional individuals were in other cohorts, that number is given in brackets. The *P*-value reported is the minimum *P*-value from the testing of the DDD dataset or the meta-analysis dataset. The subset providing the *P*-value is also listed. Mutations are considered clustered if the *P*-value proximity clustering of *de novo* mutations is less than 0.01.

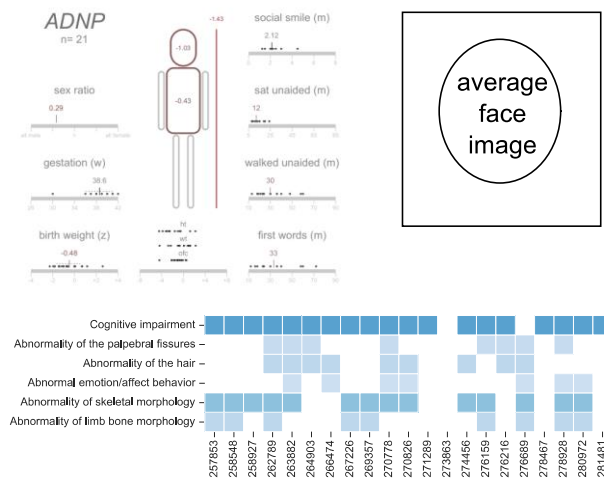
Supplementary Table 5: Counts of individuals from phenotypes used for meta-analysis, including counts of individuals with likely pathogenic *de novo* mutations in known dominant developmental disorder genes. n, number of individuals.

Disorder	Individuals (n)	With likely pathogenic (n)
DDD trios	4294	1136
Autism spectrum disorder	2780	226
Seizures	356	57
Intellectual disability	151	49

## Supplementary figures



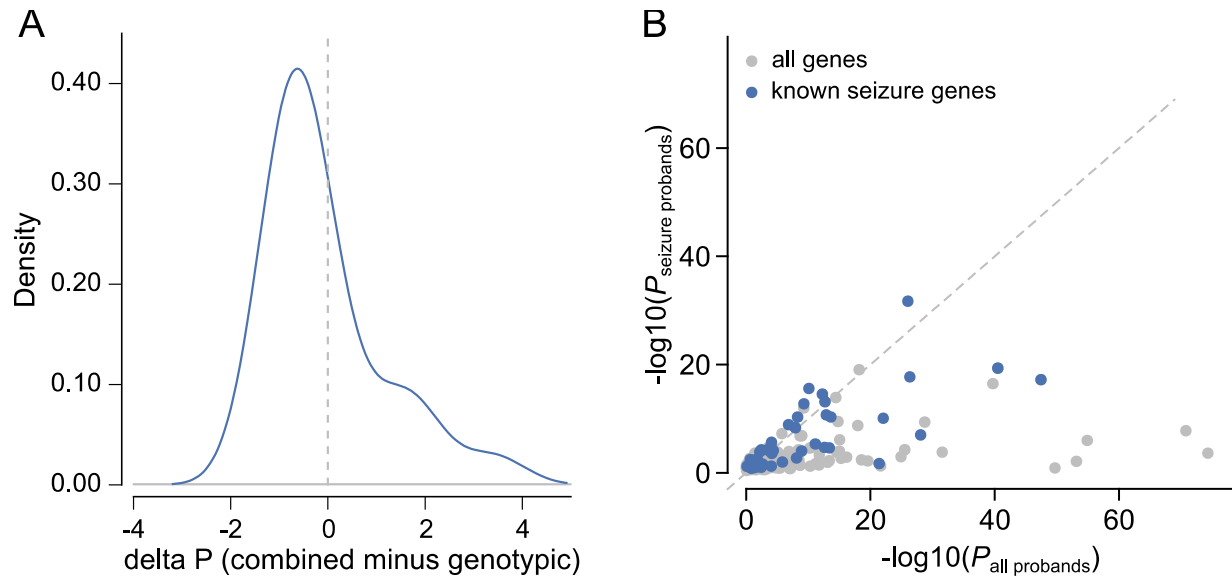
Supplementary Figure 1: Proportion of individuals with a *de novo* mutation (DNM) likely to be pathogenic. These only included individuals with protein altering or protein truncating DNMs in dominant or X-linked dominant developmental disorder (DD) associated genes, or males with DNMs in hemizygous DD-associated genes. The proportions given are for those individuals with any DNMs rather than the total number of individuals in each subset.



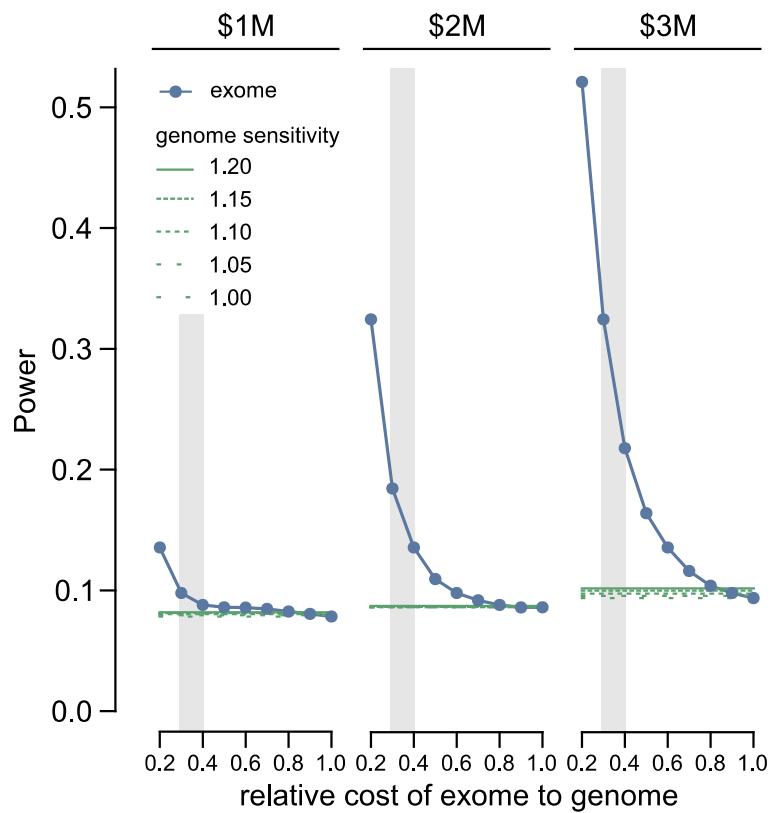
Supplementary Figure 2: Phenotypic summaries of genes exceeding genome-wide significance. Each gene subfigure has up to three parts. The first part summarises the anthropometric and developmental milestones from individuals with *de novo* mutations (DNMs) in the gene. The second part summarises the key Human Phenotype Ontology (HPO) terms for each gene. The HPO terms in the individuals were selected, including the ancestral terms. Terms that are rarer in the 4,294 individuals rank higher, adjusted by the number of individuals with DNMs who had the term. The heatmaps are shaded by the number of individuals with each term. The heatmaps exclude terms that rank lower than a descendant term (excluding more general terms if a more specific term occurred first), and terms where fewer than 25% of individuals had the term, or in genes with less than 8 individuals, terms with fewer than two individuals. The third part summarises the facial photographs from individuals with DNMs in each gene. The averaged face images are only available for selected genes, based on the availability of sufficient high-quality facial photographs of individuals for each gene.



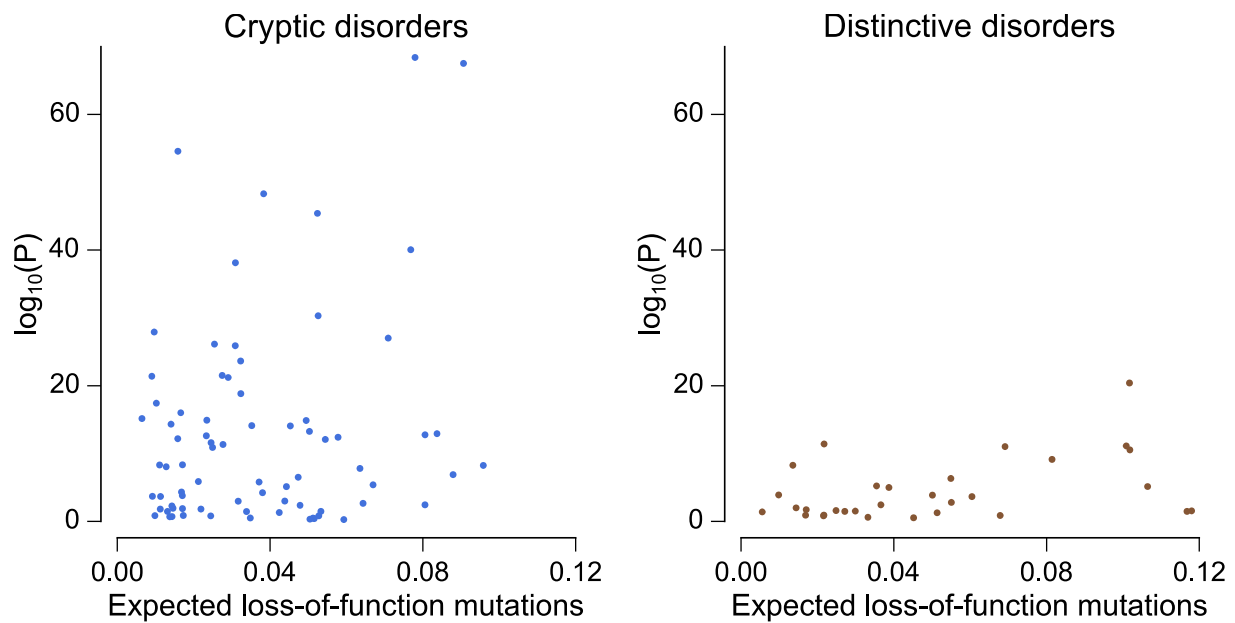




Supplementary Figure 4: Effect of clustering by phenotype on the ability to identify genomewide significant genes. A) Comparison of P-values derived from genotypic information alone versus P-values that incorporate genotypic information and phenotypic similarity. B) Comparison of P-values from tests in the complete DDD cohort versus tests in the subset with seizures. Genes that were previously linked to seizures are shaded blue.



Supplementary Figure 5: Simulated estimates of power to detect loss-of-function genes in the genome at difference cohort sizes, given fixed budgets.



Supplementary Figure 6: Neurodevelopmental genes classified by clinical recognisability were compared for the gene-wise significance versus the expected number of mutations per gene. Points are shaded by recognisability category. Genes have been separated into two plots, one plot with genes for cryptic disorders with low, mild or moderate clinical recognisability, and one plot with genes for distinctive disorders with high clinical recognisability.